Supplementary Material: On Multiset Selection with Size Constraints

Chao Qian[1], Yibo Zhang[1], Ke Tang[2], Xin Yao[2]

[1]Anhui Province Key Lab of Big Data Analysis and Application, School of Computer Science and Technology, University of Science and Technology of China, Hefei 230027, China

[2]Shenzhen Key Lab of Computational Intelligence, Department of Computer Science and Engineering, Southern University of Science and Technology, Shenzhen 518055, China

chaoqian@ustc.edu.cn, zyb233@mail.ustc.edu.cn, {tangk3, xiny}@sustc.edu.cn

Detailed Proofs

This part aims to provide some detailed proofs, which are omitted in our original paper due to space limitation.

Proof of Lemma 5. By the definition of the DR-submodularity ratio (i.e., Definition 8 in the original paper), we have

$$\begin{aligned}
\beta_f &= \min_{\boldsymbol{x}\le\boldsymbol{y}, i\in[n]} \frac{f(\boldsymbol{x}+\boldsymbol{\chi_i})-f(\boldsymbol{x})}{f(\boldsymbol{y}+\boldsymbol{\chi_i})-f(\boldsymbol{y})}\\
&\ge \min_{\boldsymbol{x}\le\boldsymbol{y}, i\in[n]} \frac{f(\boldsymbol{x}+\boldsymbol{\chi_i})-f(\boldsymbol{x})}{f(\boldsymbol{x}+(y_i-x_i)\boldsymbol{\chi_i}+\boldsymbol{\chi_i})-f(\boldsymbol{x}+(y_i-x_i)\boldsymbol{\chi_i})}\\
&= \min_{\boldsymbol{x}\in\mathbb{Z}_+^V, i\in[n], 1\le m\le c_i-x_i} \frac{f(\boldsymbol{x}+\boldsymbol{\chi_i})-f(\boldsymbol{x})}{f(\boldsymbol{x}+m\boldsymbol{\chi_i})-f(\boldsymbol{x}+(m-1)\boldsymbol{\chi_i})},
\end{aligned} \tag{1}$$

where the first inequality is by Lemma 2 since f is submodular. We then calculate $f(\boldsymbol{x}+m\boldsymbol{\chi_i})-f(\boldsymbol{x}+(m-1)\boldsymbol{\chi_i})$ for any $1\le m\le c_i-x_i$. By the definition of the objective function f (i.e., Definition 4 in the original paper), we get

$$\begin{aligned}
&f(\boldsymbol{x}+m\boldsymbol{\chi_i})-f(\boldsymbol{x}+(m-1)\boldsymbol{\chi_i})\\
&=\Big(p_i^{x_i+m}\prod_{j=1}^{m-1}(1-p_i^{(x_i+j)})\Big)\sum_{t:(v_i,t)\in E}\ \prod_{r:(v_r,t)\in E}\ \prod_{j=1}^{x_r}(1-p_r^{(j)}).
\end{aligned}$$

Note that in the above equality, $\boldsymbol{x}+m\boldsymbol{\chi_i}$ and $\boldsymbol{x}+(m-1)\boldsymbol{\chi_i}$ are different only on the budget of the source node v_i, and thus only the probabilities of activating those target nodes adjacent to v_i will be affected. By applying this equality to Eq. (1), we get

$$\begin{aligned}
\beta_f &\ge \min_{\boldsymbol{x}\in\mathbb{Z}_+^V, i\in[n], 1\le m\le c_i-x_i} \frac{p_i^{(x_i+1)}}{p_i^{(x_i+m)}\prod_{j=1}^{m-1}(1-p_i^{(x_i+j)})}\\
&\ge \min_{\boldsymbol{x}\in\mathbb{Z}_+^V, i\in[n], 1\le m\le c_i-x_i} \frac{p_i^{(x_i+1)}}{p_i^{(x_i+m)}}\\
&= \min_{i\in[n], 1\le j\le r\le c_i} \frac{p_i^{(j)}}{p_i^{(r)}}.
\end{aligned}$$

□

Proof of Lemma 6. It is easy to see that the analysis on β_f (i.e., Eq. (1)) in the proof of Lemma 5 still holds here, since it only relies on the submodularity of f. We then calculate $f(\boldsymbol{x}+m\boldsymbol{\chi_i})-f(\boldsymbol{x}+(m-1)\boldsymbol{\chi_i})$ for any $1\le m\le c_i-x_i$. By the definition of the objective function f (i.e., Definition 5 in the original paper), we get

$$\begin{aligned}
&f(\boldsymbol{x}+m\boldsymbol{\chi_i})-f(\boldsymbol{x}+(m-1)\boldsymbol{\chi_i})\\
&=\sum_{t\in T}\sum_{l}\lambda_{t,l}f_{t,l}(\boldsymbol{x}+m\boldsymbol{\chi_i})-\sum_{t\in T}\sum_{l}\lambda_{t,l}f_{t,l}(\boldsymbol{x}+(m-1)\boldsymbol{\chi_i})\\
&=\sum_{t:(v_i,t)\in E}\sum_{l}\lambda_{t,l}(f_{t,l}(\boldsymbol{x}+m\boldsymbol{\chi_i})-f_{t,l}(\boldsymbol{x}+(m-1)\boldsymbol{\chi_i})),
\end{aligned}$$

where the second equality is because for $\boldsymbol{x}+m\boldsymbol{\chi_i}$ and $\boldsymbol{x}+(m-1)\boldsymbol{\chi_i}$, only the probabilities of activating those target nodes adjacent to v_i are different. We then calculate $f_{t,l}(\boldsymbol{x}+m\boldsymbol{\chi_i})-f_{t,l}(\boldsymbol{x}+(m-1)\boldsymbol{\chi_i})$ by the definition of $f_{t,l}(\boldsymbol{x})$ (i.e., Eq. (3) in the original paper). For notational convenience, we denote $\boldsymbol{x}+m\boldsymbol{\chi_i}$ and $\boldsymbol{x}+(m-1)\boldsymbol{\chi_i}$ by $\boldsymbol{y}$ and $\boldsymbol{z}$, respectively. Then, we have

$$\begin{aligned}
&f_{t,l}(\boldsymbol{x}+m\boldsymbol{\chi_i})-f_{t,l}(\boldsymbol{x}+(m-1)\boldsymbol{\chi_i}) = f_{t,l}(\boldsymbol{y})-f_{t,l}(\boldsymbol{z})\\
&=\prod_{r:(v_r,t)\in E}\ \prod_{j=1}^{\min\{z_r,l-1\}}(1-p_r^{(j)})\prod_{j=l}^{z_r}(1-q_r^{(j)})\\
&\quad-\prod_{r:(v_r,t)\in E}\ \prod_{j=1}^{\min\{y_r,l-1\}}(1-p_r^{(j)})\prod_{j=l}^{y_r}(1-q_r^{(j)})\\
&=\prod_{r:(v_r,t)\in E}\ \prod_{j=1}^{\min\{z_r,l-1\}}(1-p_r^{(j)})\prod_{j=l}^{z_r}(1-q_r^{(j)})\\
&\quad-(1-\xi_{i,m,l})\prod_{r:(v_r,t)\in E}\ \prod_{j=1}^{\min\{z_r,l-1\}}(1-p_r^{(j)})\prod_{j=l}^{z_r}(1-q_r^{(j)})\\
&=\xi_{i,m,l}\prod_{r:(v_r,t)\in E}\ \prod_{j=1}^{\min\{z_r,l-1\}}(1-p_r^{(j)})\prod_{j=l}^{z_r}(1-q_r^{(j)})\\
&=\xi_{i,m,l}\eta_{i,m,l}\prod_{r:(v_r,t)\in E}\ \prod_{j=1}^{\min\{x_r,l-1\}}(1-p_r^{(j)})\prod_{j=l}^{x_r}(1-q_r^{(j)}),
\end{aligned}$$

where for the third equality, $\xi_{i,m,l}$ is defined as

$$\xi_{i,m,l}=\begin{cases} q_i^{(x_i+m)}, & x_i+m\ge l\\ p_i^{(x_i+m)}, & \text{otherwise}\end{cases},$$

and for the last equality, $\eta_{i,m,l}$ is the product of $m-1$ terms in the form of either $1-p_i^{(j)}$ or $1-q_i^{(j)}$ (where $x_i+1 \le j \le x_i+m-1$), since the only difference between $\boldsymbol{z} = \boldsymbol{x}+(m-1)\boldsymbol{\chi_i}$ and $\boldsymbol{x}$ is the i-th entry, i.e., $z_i = x_i+m-1$.

Let $\delta_{t,l} = \lambda_{t,l} \prod_{r:(v_r,t)\in E} \prod_{j=1}^{\min\{x_r,l-1\}} (1-p_r^{(j)}) \prod_{j=l}^{x_r} (1-q_r^{(j)})$. By applying the calculation result of $f(\boldsymbol{x}+m\boldsymbol{\chi_i}) - f(\boldsymbol{x}+(m-1)\boldsymbol{\chi_i})$ to Eq. (1), we can get

$$\beta_f \ge \min_{\boldsymbol{x}\in\mathbb{Z}_+^V, i\in[n], 1\le m\le c_i-x_i} \frac{\sum_{t:(v_i,t)\in E}\sum_l \xi_{i,1,l}\eta_{i,1,l}\delta_{t,l}}{\sum_{t:(v_i,t)\in E}\sum_l \xi_{i,m,l}\eta_{i,m,l}\delta_{t,l}}$$
$$\ge \min_{\boldsymbol{x}\in\mathbb{Z}_+^V, i\in[n], 1\le m\le c_i-x_i} \frac{\sum_{t:(v_i,t)\in E}\sum_l \xi_{i,1,l}\delta_{t,l}}{\sum_{t:(v_i,t)\in E}\sum_l \xi_{i,m,l}\delta_{t,l}},$$

where the second inequality is by $\eta_{i,1,l} = 1$ and $\eta_{i,m,l} \le 1$ for $m \ge 1$. According to the definition of $\xi_{i,m,l}$, we can divide $\sum_{t:(v_i,t)\in E}\sum_l \xi_{i,1,l}\delta_{t,l}$ and $\sum_{t:(v_i,t)\in E}\sum_l \xi_{i,m,l}\delta_{t,l}$ into three parts, respectively. That is,

$$\sum_{t:(v_i,t)\in E}\sum_l \xi_{i,1,l}\delta_{t,l} = q_i^{(x_i+1)} \sum_{t:(v_i,t)\in E}\sum_{l=1}^{x_i+1}\delta_{t,l} +$$
$$p_i^{(x_i+1)} \sum_{t:(v_i,t)\in E}\sum_{l=x_i+2}^{x_i+m}\delta_{t,l} + p_i^{(x_i+1)} \sum_{t:(v_i,t)\in E}\sum_{l>x_i+m}\delta_{t,l};$$
$$\sum_{t:(v_i,t)\in E}\sum_l \xi_{i,m,l}\delta_{t,l} = q_i^{(x_i+m)} \sum_{t:(v_i,t)\in E}\sum_{l=1}^{x_i+1}\delta_{t,l} +$$
$$q_i^{(x_i+m)} \sum_{t:(v_i,t)\in E}\sum_{l=x_i+2}^{x_i+m}\delta_{t,l} + p_i^{(x_i+m)} \sum_{t:(v_i,t)\in E}\sum_{l>x_i+m}\delta_{t,l}.$$

Note that the corresponding ratios of these three parts are $q_i^{(x_i+1)}/q_i^{(x_i+m)}$, $p_i^{(x_i+1)}/q_i^{(x_i+m)}$ and $p_i^{(x_i+1)}/p_i^{(x_i+m)}$, respectively. Since their minimum must be not larger than the ratio of the sum of the three parts, we have

$$\frac{\sum_{t:(v_i,t)\in E}\sum_l \xi_{i,1,l}\delta_{t,l}}{\sum_{t:(v_i,t)\in E}\sum_l \xi_{i,m,l}\delta_{t,l}} \ge \min\left\{\frac{q_i^{(x_i+1)}}{q_i^{(x_i+m)}}, \frac{p_i^{(x_i+1)}}{q_i^{(x_i+m)}}, \frac{p_i^{(x_i+1)}}{p_i^{(x_i+m)}}\right\}$$
$$= \min\left\{\frac{q_i^{(x_i+1)}}{q_i^{(x_i+m)}}, \frac{p_i^{(x_i+1)}}{p_i^{(x_i+m)}}\right\},$$

where the equality is by the setting $q_i^{(j)} \le p_i^{(j)}$ of the problem in Definition 5. Thus, we get

$$\beta_f \ge \min_{\boldsymbol{x}\in\mathbb{Z}_+^V, i\in[n], 1\le m\le c_i-x_i} \min\left\{\frac{q_i^{(x_i+1)}}{q_i^{(x_i+m)}}, \frac{p_i^{(x_i+1)}}{p_i^{(x_i+m)}}\right\}$$
$$= \min_{i\in[n], 1\le j\le r\le c_i} \min\left\{\frac{p_i^{(j)}}{p_i^{(r)}}, \frac{q_i^{(j)}}{q_i^{(r)}}\right\}.$$

□

For the generalized influence maximization problem as presented in Definition 1, we prove in Proposition 1 that the objective function $f(\boldsymbol{x}) = \mathbb{E}\big[|\bigcup_{l\ge1} A(X_l)|\big]$ is monotone submodular, if the fundamental propagation model Independence Cascade (Goldenberg, Libai, and Muller 2001) is used. For the Independence Cascade model as shown in Definition 2, it starts from a seed set $A_0 = X$ and uses a set A_t to record the nodes activated at time t; at time $t+1$, each inactive neighbor v_j of $v_i \in A_t$ becomes active with probability $p_{i,j}$; this process is repeated until no nodes get activated at some time.

Definition 1 (Generalized Influence Maximization)**.** *Given a directed graph $G = (V, E)$, capacities c_i ($i \in [n]$), edge probabilities $p_{i,j}$ ($(v_i, v_j) \in E$), and a budget k, it is to find a multiset $\boldsymbol{x} \in \mathbb{Z}_+^V$ such that*

$$\arg\max_{\boldsymbol{x}\le\boldsymbol{c}} \quad \mathbb{E}\big[|\bigcup_{l\ge1} A(X_l)|\big] \quad \textit{s.t.} \quad |\boldsymbol{x}| \le k,$$

where $X_l = \{v_i \mid x_i \ge l\}$ and $A(X_l)$ is the number of nodes activated by propagating from X_l.

Definition 2 (Independence Cascade)**.** *(Goldenberg, Libai, and Muller 2001) Given a directed graph $G = (V, E)$ with edge probabilities $p_{i,j}$ for any $(v_i, v_j) \in E$ and a seed set $X \subset V$, it propagates as follows:*
1. let $A_0 = X$ and $t = 0$.
2. repeat until $A_t = \emptyset$
3. for each edge (v_i, v_j) with $v_i \in A_t$ and $v_j \in V \setminus \bigcup_{r\le t} A_r$
4. v_j is added into A_{t+1} with probability $p_{i,j}$.
5. let $t = t+1$.

Proposition 1. *If the Independence Cascade propagation model is used, the objective function $f(\boldsymbol{x}) = \mathbb{E}\big[|\bigcup_{l\ge1} A(X_l)|\big]$ of generalized influence maximization is monotone and submodular.*

Proof. The monotonicity of f is trivial. We are to prove its submodularity. According to Lemma 2 in the original paper, we only need to prove that for any $\boldsymbol{x} \le \boldsymbol{y}$ and $i \in [n]$ with $x_i = y_i$,

$$f(\boldsymbol{x}+\boldsymbol{\chi_i}) - f(\boldsymbol{x}) \ge f(\boldsymbol{y}+\boldsymbol{\chi_i}) - f(\boldsymbol{y}).$$

Let $X_l = \{v_j \mid x_j \ge l\}$. According to the definition of the objective function f, we get

$$f(\boldsymbol{x}+\boldsymbol{\chi_i}) - f(\boldsymbol{x})$$
$$= \mathbb{E}\big[|A(X_{x_i+1}\cup\{v_i\}) \cup \bigcup_{l\ne x_i+1} A(X_l)|\big]$$
$$- \mathbb{E}\big[|A(X_{x_i+1}) \cup \bigcup_{l\ne x_i+1} A(X_l)|\big].$$

Let $G' = (V, E')$ denote a subgraph of $G = (V, E)$, which is generated by preserving each edge $(v_i, v_j) \in E$ with probability $p_{i,j}$. Then, the set of nodes reachable from X_l in G' actually corresponds to $A(X_l)$. Note that G' is random. For each fixed G', it is easy to see that $A(X_{x_i+1}) \subseteq A(X_{x_i+1}\cup\{v_i\})$, since $X_{x_i+1} \subseteq X_{x_i+1}\cup\{v_i\}$. Let $S = \bigcup_{l\ne x_i+1} A(X_l)$. Thus, we have

$$f(\boldsymbol{x}+\boldsymbol{\chi_i}) - f(\boldsymbol{x}) = \mathbb{E}\big[|A(X_{x_i+1}\cup\{v_i\}) \setminus A(X_{x_i+1}) \setminus S|\big].$$

Let $Y_l = \{v_j \mid y_j \geq l\}$ and $T = \bigcup_{l \neq y_i+1} A(Y_l)$. We can similarly get

$$f(\boldsymbol{y}+\boldsymbol{\chi_i})-f(\boldsymbol{y})=\mathbb{E}\big[|A(Y_{y_i+1}\cup\{v_i\}) \setminus A(Y_{y_i+1}) \setminus T|\big].$$

Since $x_i = y_i$, T is actually $\bigcup_{l \neq x_i+1} A(Y_l)$, and

$$f(\boldsymbol{y}+\boldsymbol{\chi_i})-f(\boldsymbol{y})=\mathbb{E}\big[|A(Y_{x_i+1}\cup\{v_i\}) \setminus A(Y_{x_i+1}) \setminus T|\big].$$

Note that $X_l \subseteq Y_l$, since $\boldsymbol{x} \leq \boldsymbol{y}$. Thus, for any l, it holds that $A(X_l) \subseteq A(Y_l)$ for each fixed subgraph. This implies that $S \subseteq T$. Furthermore, for each fixed subgraph, $A(Y_{x_i+1} \cup \{v_i\}) \setminus A(Y_{x_i+1}) \subseteq A(X_{x_i+1} \cup \{v_i\}) \setminus A(X_{x_i+1})$, since $A(X_{x_i+1}) \subseteq A(Y_{x_i+1})$. Thus, we can get

$$(f(\boldsymbol{x}+\boldsymbol{\chi_i})-f(\boldsymbol{x})) - (f(\boldsymbol{y}+\boldsymbol{\chi_i})-f(\boldsymbol{y})) \geq 0.$$

Thus, the proposition holds. □

References

Goldenberg, J.; Libai, B.; and Muller, E. 2001. Talk of the network: A complex systems look at the underlying process of word-of-mouth. *Marketing Letters* 12(3):211–223.